

(19) World Intellectual Property
Organization
International Bureau



(43) International Publication Date
2 June 2005 (02.06.2005)

PCT

(10) International Publication Number
WO 2005/050472 A2

(51) International Patent Classification⁷: **G06F 17/20**

[NL/NL]; Groenewoudseweg 1, NL-5621 BA Eindhoven (NL).

(21) International Application Number:
PCT/IB2004/052404

(22) International Filing Date:
12 November 2004 (12.11.2004)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
03104315.1 21 November 2003 (21.11.2003) EP

(71) Applicant (for DE only): **PHILIPS INTELLECTUAL PROPERTY & STANDARDS GMBH** [DE/DE]; Stein-
damm 94, 20099 Hamburg (DE).

(72) Inventors; and

(75) Inventors/Applicants (for US only): **PETERS, Jochen** [DE/DE]; c/o Philips Intellectual Property & Standards GmbH, Weisshausstr. 2, 52066 Aachen (DE). **MEYER, Carsten** [DE/DE]; c/o Philips Intellectual Property & Standards GmbH, Weisshausstr. 2, 52066 Aachen (DE). **KLAKOW, Dietrich** [DE/DE]; c/o Philips Intellectual Property & Standards GmbH, Weisshausstr. 2, 52066 Aachen (DE). **MATUSOV, Evgeny** [RU/DE]; c/o Philips Intellectual Property & Standards GmbH, Weisshausstr. 2, 52066 Aachen (DE).

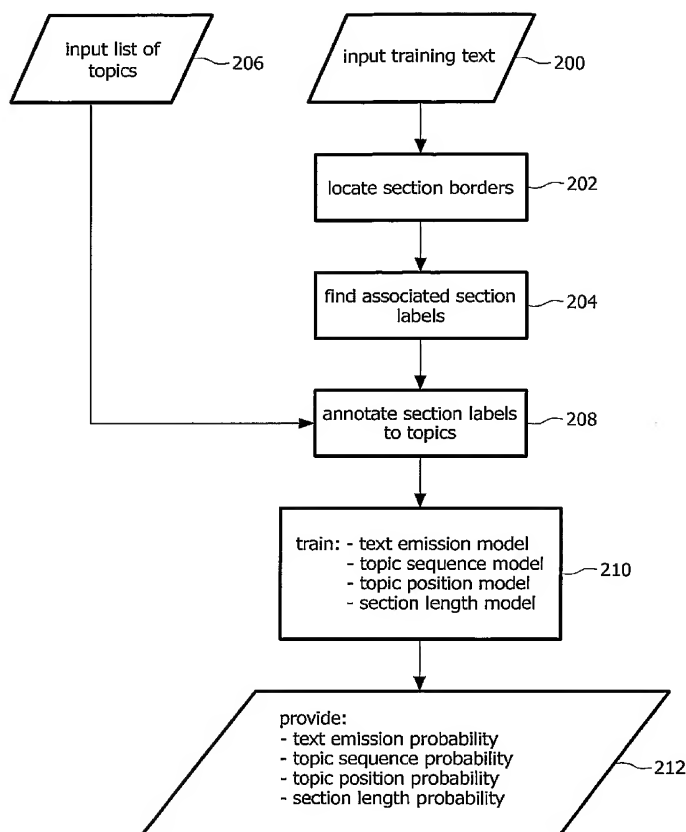
(74) Agents: **MEYER, Michael** et al.; Philips Intellectual Property & Standards GmbH, Weisshausstr. 2, 52066 Aachen (DE).

(71) Applicant (for all designated States except DE, US): **KONINKLIJKE PHILIPS ELECTRONICS N. V.**

(81) Designated States (unless otherwise indicated, for every kind of national protection available): AE, AG, AL, AM,

[Continued on next page]

(54) Title: TEXT SEGMENTATION AND TOPIC ANNOTATION FOR DOCUMENT STRUCTURING



(57) Abstract: The invention relates to a method, a computer program product and a computer system for structuring an unstructured text by making use of statistical models trained on annotated training data. Each section of text in which the text is segmented is further assigned to a topic which is associated to a set of labels. The statistical models for the segmentation of the text and for the assignment of a topic and its associated labels to a section of text explicitly accounts for: correlations between a section of text and a topic, a topic transition between sections, a topic position within the document and a (topic-dependent) section length. Hence structural information of the training data is exploited in order to perform segmentation and annotation of unknown text.

WO 2005/050472 A2



AT, AU, AZ, BA, BB, BG, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NA, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RU, SC, SD, SE, SG, SK, SL, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, YU, ZA, ZM, ZW.

(84) Designated States (*unless otherwise indicated, for every kind of regional protection available*): ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM),

European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IS, IT, LU, MC, NL, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:

— *without international search report and to be republished upon receipt of that report*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.